Layer-wise relevance propagation for backbone identification in discrete fracture networks

Recognition and classification of typical energy consumption profiles in buildings with non-intrusive learning approach

Marco Savino Piscitelli[a], Silvio Brandi [a] and Alfonso Capozzoli[a,*]

[a] Dipartimento Energia "Galileo Ferraris", Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
* Corresponding author: Tel: +39-011-090-4413, fax: +39-011-090-4499, e-mail: alfonso.capozzoli@polito.it

Abstract:

The progressive spread of Advanced Metering Infrastructure (AMI) has enabled the collection of a huge amount of building related-data which can be exploited by both energy suppliers and users to gain insight on energy consumption patterns. In this context, data analytics-based methodologies play a key role for performing advanced characterization, benchmarking and classification of buildings according to their typical energy use in the time domain. Traditionally, energy customers are classified according to their building end-use category. However, buildings belonging to the same category can exhibit different energy patterns making ineffective this kind of a-priori categorization. For this reason, load profiling frameworks have been developed in the last decade to identify homogenous groups of buildings with similar daily energy profiles. The present study proposes a non-intrusive customer classification process which does not make use of in-field load monitoring data for the classification of unknown customers. The classification process is developed by analysing hourly energy consumption data of 114 electrical customers of an Italian Energy Provider. The representative daily load profiles are grouped using the "Follow the Leader" clustering algorithm and a globally optimal decision tree is employed to build a supervised classification model. This model is also compared to a baseline recursive partitioning tree leading to an increasing of accuracy of about 6%. The predictive attributes are gathered from monthly energy bills and from additional information on customers' habits collected by means of phone survey. Eventually, the procedure exploits energy bill data also for estimating the magnitude of typical load profiles.

Highlights:

- A stock of buildings is analyzed to discover typical energy consumption profiles;
- The daily load profiles are grouped with a "Follow the Leader" clustering algorithm;
- A globally optimal decision tree is employed to develop a customer classifier;
- The proposed classifier performs better than the baseline model of about 6%;
- The classifier makes use of non-intrusive attributes gathered from energy bills

Keywords: building energy consumption, customer classification, load profiling, data mining, energy benchmarking.

---

1. Introduction

The progressive introduction of Advanced Metering Infrastructure (AMI) in the last years has enabled the collection of a huge amount of building energy consumption data [1,2]. In this context, data analytics methodologies can be exploited by energy suppliers to gain insight on energy consumption patterns for a vast number of buildings [3]. A significant amount of research has been conducted in the field of building characterization using measured meter data [4]. Two comprehensive reviews on the use of unsupervised

learning and data analytics techniques for the analysis of building operational data were recently published discussing relevant applications in this area [5,6].

The most promising applications of data analytics in energy buildings deal with the prediction of the energy demand of buildings [7], the optimization of building system operation [8], the detection and commissioning of operational status and failures of building equipment [9], the assessment of the impact of user on building energy consumption [10,11], the energy benchmarking analysis [4], the charachterisation of building energy demand [12,13]. The latter field of research often deals with the exploitation of various extracted temporal features from smart meter data [14] (e.g., load shape features, weather-dependency features, load pattern specificity, load diversity, long and medium-term volatility) for the segmentation and classification of large stock of buildings according to their energy behavior. In fact, mining the energy consumption patterns of buildings in large stocks not only provide more robust energy benchmarks [15,16] but can also support the development of energy management initiatives and demand response programs [17] targeted to specific segments of users [18].

Traditionally, energy customers are segmented and classified according to their building end-use category as residential, industrial, commercial and so on. However, in many cases, customers belonging to the same category can exhibit significantly different patterns in their energy consumption [15,19]. In such cases, benchmarking methods related to the energy use intensity (e.g., kWh/m^3y) of the building are not able to fully characterise the energy behaviour of a customer over time. On the contrary, knowledge extracted from energy consumption time series (i.e., load profiling analysis) contains information on how and when building energy use changes during the day for various end uses such as appliances, lighting, ventilation, heating and cooling [16,20].

A number of load profiling frameworks have been developed in the literature to deal with data coming from multiple buildings usually with the aim to identify, through unsupervised analysis, homogenous groups of typical daily load curves (i.e., customer classification) characterised by similar shapes and/or magnitude [3,21]. An in-depth knowledge of building typical load profiles could help managers in addressing some emerging challenges in the energy and buildings research field, such as energy consumption forecasting [22], anomaly detection and diagnosis through the introduction of robust analytics framework [23] and the evaluation of infrequent and unexpected daily energy patterns [24,25]. Moreover, the recognition of energy demand profiles enables the development of robust demand side management (DSM) strategies [26,27]. Rhodes et al. in [28] stated that load profiling of residential customers could serve as a starting point for utilities looking to reduce electricity use during peak times by developing policies that target load shifting. Eventually, in the two-way paradigm of smart grid, load profiling is particularly beneficial for both energy providers and users that are involved in Demand Response (DR) programs [17,29]. In the current competitive energy retail market, DR programs are designed to be attractive for the consumers and at the same time profitable for the retailers. In incentive-based programs, knowledge of customers' macro-behavior in energy consumption allows the distribution companies to better manage the grid operation [30] and the interactions between energy consumption and production [17,31] (e.g., indirectly switching certain electric appliances at certain times). The modification of a load profile allows to flat the demand profile or in some cases to follow the generation pattern for grid stability purpose [32]. For example, virtual thermal storage, through the modification of load profiles of a group of buildings served by a district heating network represents an effective way to increase the share of heat from cogeneration and renewable sources [33]. In particular identify consumers who exhibit more-variable load patterns on normal days is essential as they could be able to change their loads more significantly when involved in demand response programs [34]. In that perspective energy retailers can take advantage from that knowledge in the design stage of dynamic pricing plans. According to the different customer groups in the portfolio, different energy tariffs can be set for each typical curve in order to maximize the relative profit [35,36]. For instance, in [37] Chicco et al. demonstrated how a data-driven customer classification process could be used to modify existing energy tariffs by fixing rate coefficients for each customer class.

Also the customer side is experiencing a revolution in the smart grid environment in terms of demand management opportunities. In fact, thanks to the spread of electrical/thermal energy storages, renewable energy systems and data analytics-based technologies in buildings [38], user's energy demand is becoming more and

more flexible [31,32]. Energy managers can implement, in an easier way, strategies aimed at modifying building energy use to obtain targeted changes in electrical/thermal load profile [32]. In this way, customers can change their load profiles (e.g., consuming less energy during peak hours or shifting the energy use to off-peak hours) in response of variations of energy price over time [19] (i.e., price-based programs) leveraging on energy flexibility and fully exploiting building potential in the energy management [39]. Benchmarking the energy usage in the time domain, through load profiling, is then crucial also for the impact assessment of DMSs and DR initiatives [40,41]. The information about shape and magnitude of electrical power consumption patterns can reveal useful knowledge [42] about building energy flexibility potential and/or in some cases the presence of multiple typical patterns (e.g., seasonality, intra-week variation)[28].

From the design point of view, the in-depth characterization of the energy demand makes it possible to better address the current transition from large centralized generation plants to multi-energy distributed ones that are capable to provide, from different sources, energy at a small scale (e.g., neighborhood) when it is needed [32]. In fact, the lack of knowledge about building energy use patterns currently represents the main barrier for fully exploiting the benefits of energy management also at micro grid level.

For the sake of completeness, an overview on the methodological process and methods employed for addressing the customer classification task is provided in the next section of the paper.

1.1. Overview of customer classification process through load profiling analysis

In the literature, the customer classification problem has been widely discussed by several researchers. Overviews on data mining based methodologies for customer classification are provided in [3,43,44]. Typically, this task unfolds through four main methodological stages: i) identification of *n* classes of customers with similar energy consumption profiles; ii) definition of the reference load pattern for each class; iii) enrichment of the database with predictive attributes; iv) development of a supervised classification model.

The first stage of the process, in most of the cases, makes use of unsupervised data mining or machine learning techniques to identify homogenous groups of customers based on their electrical/thermal daily load profiles [43]. To address that task several algorithms were proposed in the literature and tested on different case studies (e.g., from low voltage to high voltage electric customers).

According to Panapakidis et al [45] the methods used for the identification of homogenous load profile groups can be categorized as partitional clustering algorithms (e.g., k-means), fuzzy clustering algorithms (e.g., Fuzzy C-means), hierarchical clustering algorithms, neural network based clustering (e.g., self-organizing maps) and algorithms that not belong to the previous categories (e.g., support vector clustering). The k-means algorithm was used with success for the classification of industrial [46] or domestic [47] electricity customers. Fernandes et al. used the Fuzzy C-means for the segmentation of residential gas consumers [48]. In [49] a customer classification process was performed by using a hierarchical clustering process, while Figueiredo et al. characterised the energy consumers by means of a self-organizing maps [50]. Moreover in [51] a support vector clustering process was adopted to segment electrical load patterns.

Despite their proven effectiveness, the robustness of such unsupervised methods is strictly dependent from various factors such as the aggregation algorithm (e.g., complete, single linkage in hierarchical clustering) [43], the dissimilarity distance measure between profiles [52,53], the data normalization technique [3] and number of clusters (i.e., customer groups). Due to such degrees of freedom in the clustering problem formulation, several adequacy indices (based on the measure of inter-cluster similarity and intra-cluster dissimilarity) have been proposed in the literature in order to assess the quality of clustering results [43].

In [43,45] the most popular indices were reviewed such as mean index adequacy (MIA), Clustering Dispersion Indicator (CDI), Scatter Index (SI), Silhouette index, Variance Ratio Criterion (VRC) and Davies-Bouldin Index (DBI). The use of adequacy indices makes it possible to partially supervise the process suggesting the most suitable number of customer groups to be assumed in the clustering analysis.

The outcome of that stage is then the identification of a number of customer classes (buildings with similar energy consumption profiles), for which the reference load pattern can be calculated as the centroid or medoid of the profiles grouped together. Subsequently, the customer class label is encoded as a categorical variable to

be predicted through a supervised classification model. To this purpose the load profile database is usually enriched with additional attributes (categorical and/or numerical) to be considered in the classification as predictive variables.
These attributes can be defined as a-priori or based on in-field measurement campaign according to [37]. A-priori indicators are related to the customers' energy contracts and the type of commercial activity and are generally used by energy providers to preliminary characterize their clients. These indicators are static and do not exhibit sensitivity to load profile shape and magnitude [37]. Indeed, if they are used as unique predictors they cannot provide a comprehensive characterization of the energy use of customers in the time domain [3]. For this reason, indicators extracted from in-field measurement campaigns are employed in order to ensure a higher accuracy of the supervised classification model. These indicators deal with specific features of the load profile shapes and are calculated for each customers' reference load pattern.
These indicators (in the (0,1) range) are capable to capture the normalized variability in daily load profiles, and hourly/sub-hourly load shares with respect to specific reference values (mean, max, min, standard deviation) in different daily periods (e.g. night, lunch time) [37,54].

Once the predictive attributes are selected, the customer classification process goes through the development of a supervised classification model. The classification task aims at assigning unknown customers into pre-identified classes. Decision trees (e.g. C4.5, C5.0, CART) have been often used in the literature to accomplish that task due to their capability in handling both categorical and numerical variables and the high readability of their output in form of decision rules [55,56]. In [57] Ramos et al. used C5.0 algorithm for classifying a portfolio of about 1000 medium voltage customers in groups identified though a clustering analysis. Also in [50] Figueiredo et al. employed the C5.0 algorithm for customer classification purpose. In particular, a different consumer characterization is obtained for each load conditions considered. As a reference for winter working days and weekends the overall classification accuracy is close to 80% leveraging on a set of about 30 decision rules.

Although the clustering phase is well investigated in the literature, little focus has been devoted to classification phase and in particular to the nature of the predictive attributes. As previously explained, in most of cases the classification attributes are directly extracted from the load curves as done in [54,58]. These variables show an excellent explanatory potential; however, they can be computed only through an intrusive approach. This is usually unfeasible since energy retailers not always have at their disposal such information when dealing with a new customer. In response to this gap, in this paper a customer classification framework relying only on variables obtained through non-intrusive approach has been developed. The proposed framework is conceived to be employed by energy retailers and demand response operators to identify representative groups of customers in heterogeneous stocks of buildings.

The representative load profiles are grouped with a "Follow the Leader" clustering algorithm [37,49]. In the post-clustering phase, a globally-optimal decision tree [59] is employed to build a supervised classification model and compared against a traditional recursive partitioning decision tree. The non-intrusive predictive attributes are extracted from monthly energy bills of each customer and from additional information about customers' habits collected by means of phone survey.

1.2. Contribution of the paper

The present paper focuses on the analysis of electrical load patterns of a stock of industrial and commercial buildings. The entire process relies on the application of data mining based algorithms in order to develop a customer classification tool capable to estimate for an unknown customer its most probable reference load profiles. The main challenge is to develop a non-intrusive classification tool that does not take into account attributes based on in-field load monitoring as input variables.
On the basis of the literature review on customer classification presented in section 1.1, the main issues that this paper intends to address are the following:

- The most of the analytical effort presented in literature has been devoted to the pattern recognition phase (i.e., clustering phase) often neglecting the development of classifiers capable to estimate, for an unknown customer, its most probable cluster label and representative profiles;

- When a classification model is developed, in most of the cases the input attributes are gathered from in-field energy monitoring campaigns. It means that such a classifier can be used by an energy provider/retailer only for classifying customers whose energy consumption profiles are already available;
- The output of the customer classification process mainly consists in estimating normalised reference shapes of load profiles (e.g, (0,1) range) without providing any information about their magnitude;
- In most of the applications only one reference load pattern per customer is considered for the subsequent clustering analysis. This assumption while allowing the dataset to be reduced, in some cases can costraint the exploration of different load conditions (e.g., seasonal patterns).

In that perspective, the present study aims at conceiving, developing and testing a methodological framework that contributes in facing the aforementioned issues in a robust way as possible. A classification tool which makes use of few input variables collected through a non-intrusive approach was introduced. It allows typical reference patterns (including their shapes and magnitudes) to be identified by an energy provider in the very early stage of a customer engagement. As a consequence, more effective energy management strategies can be implemented for different customer groups by exploiting easy-to-collect and non-intrusive data and information (e.g., billing data, working time).

The rest of the paper is organised as follows. Section 2 provides a description of the analysed dataset. Section 3 presents the methodology adopted for non-intrusive customer classification process. Section 4 briefly describes the methods and algorithms used to perform the analysis. Section 5 presents the results obtained for the analysed case study. The last two sections discuss the results and contain the concluding remarks of the study.

2. Description of the dataset

The customer classification process has been developed starting from the monitored data of 114 electrical customers of an Italian Energy Provider (eVISO s.r.l.). The buildings are located in Piedmont (North-Western region of Italy) and are characterized by similar climate conditions. 17 customer typologies (i.e., building end uses) were taken into account in the analysis. In particular, from *Figure 1* it can be inferred that the majority of the analysed buildings are manufacturing industries (i.e., metal-working, wood-working, stone-working).

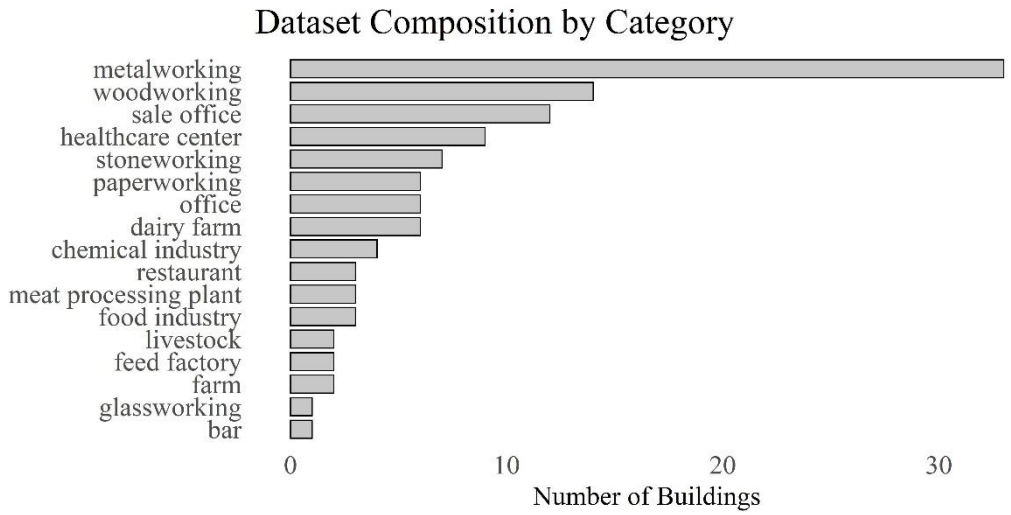

Figure 1 – Number of customers for each category

The analysed data consists of three different datasets:

- **Electrical power dataset**: it includes at least 4 months of measured hourly power demand of the 114 customers from “00:00:00 2014-01-01” to “23:00:00 2017-01-31”;

- **Energy bills dataset**: it includes the monthly billing information for the 114 customers;
- **Additional info dataset**: it includes features of the 114 customers such as customer typology and working time.

Electrical power data were collected by means of smart meters installed by the energy provider while monthly billing data and additional information were retrieved through energy bills and short phone surveys. In the present study data were analysed and presented in anonymous form due to privacy issues related to the customer's portfolio of the energy provider.

Timestep	Building 1 [kW]	Building 2 [kW]	…	Building N [kW]
2016-01-01 00:00:00	5.0	15.5		5.3
2016-01-01 01:00:00	4.0	17.6		4.7
.				
.				
.				
2016-12-31				

Building	Month	ToU F1 [kWh]	ToU F2 [kWh]	ToU F3 [kWh]
1	January 2016	1000.0	575.5	545.3
1	February 2016	12250.0	6070.6	4530.7
.	.	.	.	.
.				
.				
N				

Building	Category	Opening time	Closing time	Lunchtime duration [hours]	…
1	Office	06:00	19:00	1	
2	Metalworking	08:00	18:00	1.5	
	.	.	.	.	…
	.	.	.	.	
	.	.	.	.	
N	Woodworking	4:00	22:00	2	

Figure 2 - Example of raw data structure

As an example, Figure 2 shows an extraction of records from the available dataset in order to provide an understanding of raw data structure. The year 2016 was selected as reference period for conducting the analysis. During 2016 the *electrical power* and *energy bills* datasets present the minimum amount of missing data and all the additional info were available. In Figure 3 are also shown the box plots of the average electrical power demand in the three time slots related to different Italian electrical energy tariffs (F1, F2, F3) for the buildings in the same category. The high diversity of the sample, in terms of building typology and energy consumption, represents an asset in a customer classification process in the perspective of extracting knowledge as generalizable as possible.

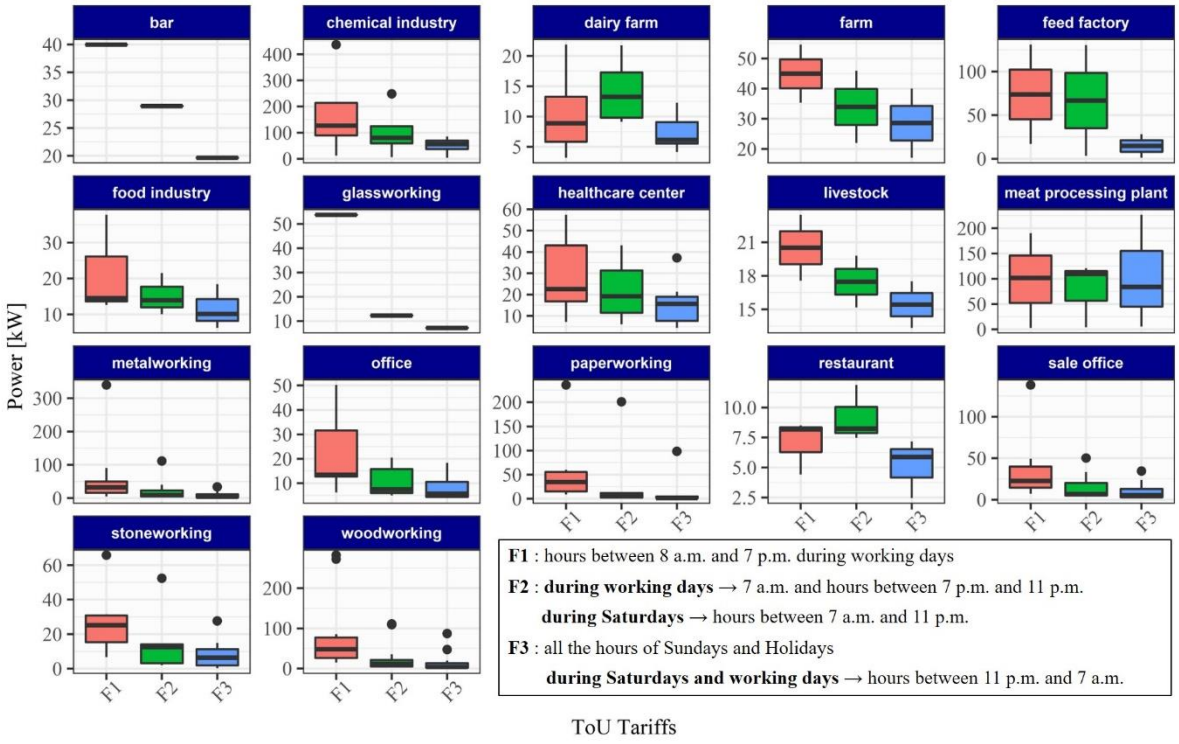

Figure 3 - Box plots of the average electrical power demand in the three time slots related to different italian electrical energy tariffs (F1, F2, F3) for the buildings in the same category

3. Methodological framework

In this section the methodological framework is presented with the aim of discussing each analytical stage of the process. The methodology relies on the application of a clustering algorithm coupled with a decision tree, to perform a robust classification of a number of electrical industrial and commercial customers. The whole process has been developed and tested on the dataset previously described in section 2. The general framework unfolds over four different stages, as shown in *Figure 4*.

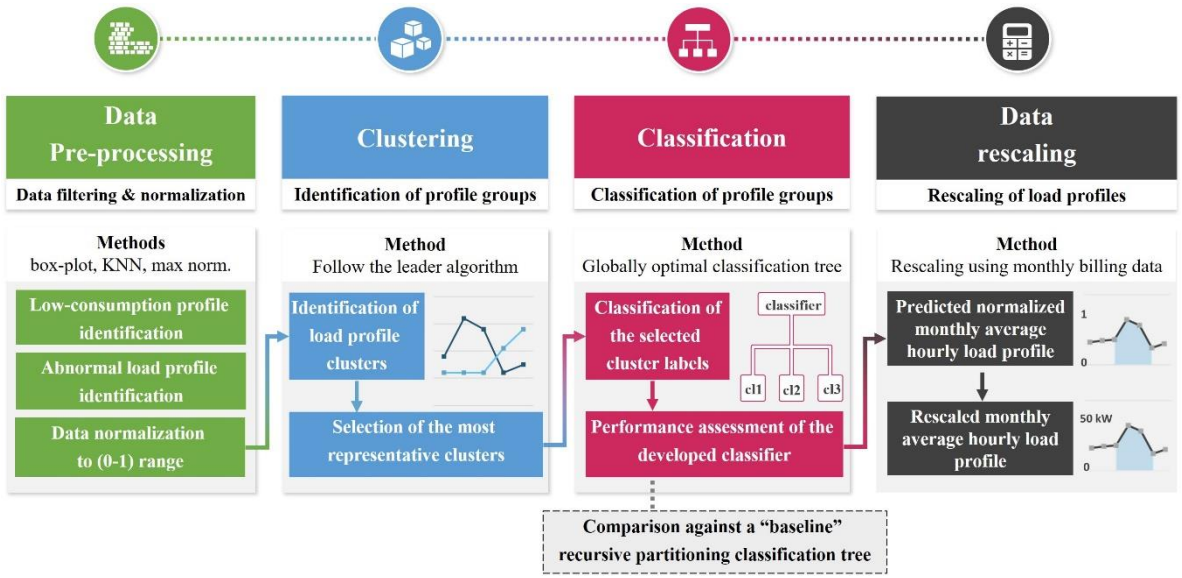

Figure 4 - General methodological framework of the analysis

Data pre-processing: the first stage is aimed at filtering and preparing the data. Data pre-processing is a mandatory task for any analytical process applied to data collected by means of smart meters. At this stage for each building the time series of hourly energy consumption was chunked into subsequences of daily length. The analysis was performed only on working days, filtering out the load profiles of weekends and holidays. The load profiles of the working days were then analysed in order to identify missing values and punctual outliers that were removed and replaced. Furthermore, all the days characterised by very low or infrequent variation in electrical load over time were removed. Finally, the remaining daily load profiles were averaged for each month and normalised in the range (0,1) resulting in 1249 Normalized Monthly Reference Load Profiles (NMRLPs). A detailed description of data pre-processing is provided in section 4.1.

Clustering: the second stage of the analysis is aimed at grouping similar NMRLPs in clusters which are representative of specific energy consumption patterns. The unsupervised segmentation was performed by means of "Follow the Leader" clustering algorithm [37,49] using the Euclidian distance as dissimilarity measure. Details on the clustering method are provided in section 4.2.

Classification: the group of clusters evaluated in the previous stage were analysed and described, and the labels of the most representative ones were used as target variables in a classification process. More in detail, a proposed model consisting in a globally optimal decision tree [55,59] was compared with a baseline model consisting in a recursive partitioning classification tree algorithm. The proposed model makes use of stochastic optimisation methods (i.e., evolutionary algorithms) that can lead to much more accurate classification than locally optimal decision trees [59]. Both the classification models (i.e., baseline and proposed) were developed using the cluster labels as target variable, and additional building features as input variables. The classifiers are able to predict, for a new customer, the most probable NMRLP on monthly basis only using a-priori knowledge (e.g., occupant arrival and exit time) and billing data. As a consequence, an energy provider may be able to easily estimate, for a new electrical customer, the monthly average hourly load profile based on the membership to a customer class previously identified. Details on the classification algorithms are provided in section 4.3.

Data rescaling: the final stage of the process consists in the rescaling of NMRLPs. In fact, after the prediction of the NMRLP for a new customer, it becomes particularly desirable the evaluation of the magnitude associated to these normalized profiles. To address this task only historical billing data were used as shown in *Figure 5*.

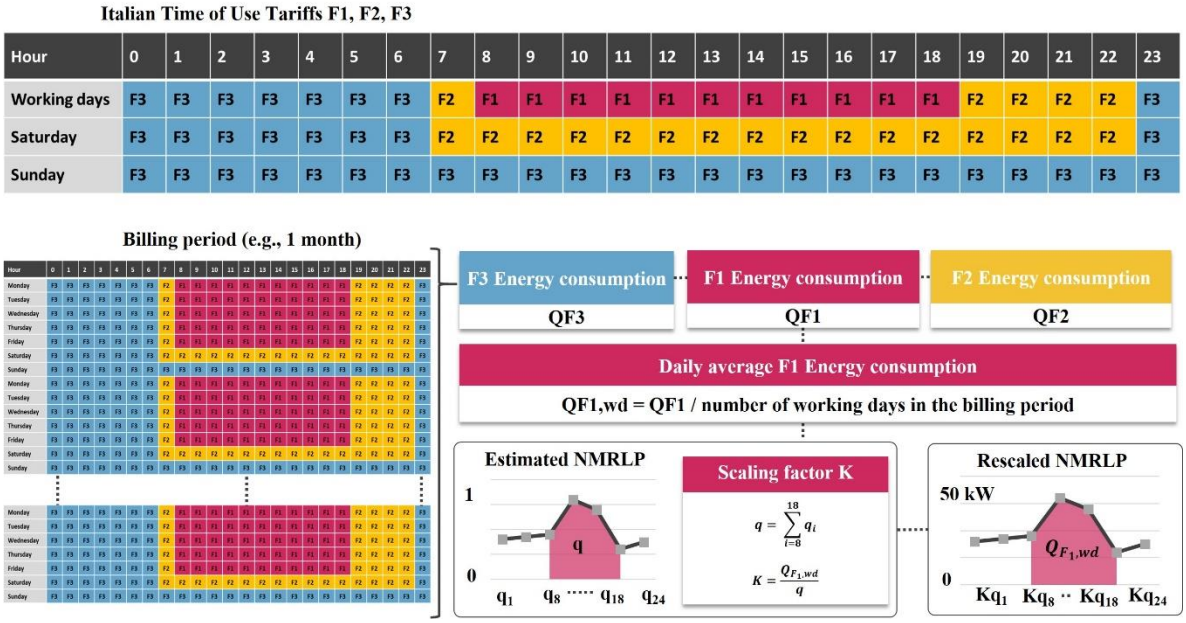

Figure 5 - Methodological process for the rescaling of the Normalized Monthly Reference Load Profiles (NMRLPs)

In Italy, from electrical energy bills, it is possible to associate energy consumption data to hours with specific Time of Use (ToU) tariffs. The Italian ToU tariffs consist of three different daily time slots (Figure 5):

- F1 time slot (peak hours): it includes hours between 8 a.m. and 7 p.m. during working days;
- F2 time slot (off-peak hours): during working days this slot includes one hour in the morning (7 a.m.) and hours between 7 p.m. and 11 p.m. During Saturdays it includes hours between 7 a.m. and 11 p.m.;
- F3 time slot (off-peak hours) which comprises the remaining hours not included in the F1 and F2 time slots (i.e., Sundays, Holidays and night hours between 11 p.m. and 7a.m.).

In the present study only working days were analysed for computing NMRLPs for each customer. For this reason, in order to rescale these normalized load profiles, only the energy consumption referred to the F1 slot during the billing period was considered, since the other slots are also included in the weekends and holidays.

Assuming a monthly billing period, the total energy consumption in the time slot F1 for that period ($\boldsymbol{Q_{F_1}}$) is divided for the number of working days to calculate the daily average energy consumption $\boldsymbol{Q_{F_1,wd}}$ expressed in kWh. The scaling factor K is then calculated as follows:

$$\boldsymbol{K = \frac{Q_{F_1,wd}}{q}} \quad (1)$$

Where q is the normalized daily average energy consumption of the estimated NMRLP during the F1 time slot (i.e., 8 a.m. – 18 p.m.) calculated as follows:

$$\boldsymbol{q = \sum_{i=8}^{18} q_i * T} \quad (2)$$

Where $\boldsymbol{q_i}$ is the i-*th* normalized average power of the NMRLP and T is the timestep of the load profile expressed in hours. After the evaluation of the scaling factor, each $\boldsymbol{q_i}$ of the NMRLP was multiplied by K. Assuming that K is calculated starting from an average energy balance on about the 50% of the hours of a day (F1 time slot), it can be considered a reliable and representative scale factor for an entire working day. For this reason, the factor K is then used also to rescale $\boldsymbol{q_i}$ not included in the F1 time slot.

Following this framework, the rescaling process has been proved to be straightforward and robust. The entire process was tested using a sampling composed by 13 customers, for which one-year of hourly data were available. The approximation error referred to classification and rescaling of NMRLPs has been evaluated in the testing phase. Based on the obtained results, the proposed methodology is a very useful and ready-to-implement tool that can be generalised for different kinds of buildings for customer classification purpose.

4. Methods and algorithms of analysis

In this section, a brief theoretical description is given for the aforementioned methods and algorithms used to perform the analysis of the reference load profiles. The analytic methods are also discussed with the aim of better specifying the advantages they offer in relation to the objectives of the present work in the field of energy customer classification. Furthermore, details on data pre-processing are provided.

4.1 Data filtering and normalization methods

The time series of energy consumption for each building was chunked into daily sub-sequences. After the segmentation, only load profiles of working days were taken into account.

Punctual outliers were removed from daily load profiles and replaced through linear interpolation. Furthermore, also outliers at daily energy trend level were detected and removed. These profiles were characterised by very low or infrequent variation in energy demand over time.

The first type of abnormal patterns is represented by days during which the electrical load was significantly lower than the other working days. These days may include holidays or days that were not correctly identified and labelled as non-working days. The identification process of these profiles was conducted separately for each customer and for each month. For each daily load profile, the daily power demand standard deviation was

calculated. In this way, through a box plot analysis for each customer and for each month, the low variation profiles were identified according to the following equations:

$$OUT_{SD} = Q1_{SD} - 1.5 \cdot IQR_{SD} \quad (3)$$

$$IQR_{SD} = Q3_{SD} - Q1_{SD} \quad (4)$$

Where Q1 and Q3 are the first and third quartile of the frequency distribution of the standard deviation of daily power demand respectively, and IQR represents the interquartile range. All the profiles labelled as OUT_{SD} were the lower outliers of the distribution and were removed from the set of data.

The second type of abnormal patterns were the days during which the electrical power demand was significantly different, in terms of magnitude and shape, from the other working days. The identification of such profiles was carried out separately for each month and for each customer in the dataset. To this purpose, the k-Nearest-Neighbours (KNN) algorithm was employed.

The algorithm computes the distance matrix between all the elements (i.e., daily load profiles) in a specific month and identifies for each profile the set of its K nearest neighbours. The number of K neighbours and the distance metric are set by the analyst. In this case study K is assumed equal to 4 and the distance metric adopted is the Euclidean distance computed as follows:

$$d = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (5)$$

The algorithm returns for each element the distance values of its 4 nearest neighbours. These 4 values were averaged into one single value and its frequency distribution among the months was computed. The outliers of these distributions represent the daily load profiles that significantly differ from their nearest neighbours and are identified according to the following equations:

$$OUT_{KNN} = Q3_{KNN} + 1.5 \cdot IQR_{KNN} \quad (6)$$

$$IQR_{KNN} = Q3_{KNN} - Q1_{KNN} \quad (7)$$

Where Q1 and Q3 are the first and third quartile of the frequency distribution of the average distance of each profile from its neighbours respectively, and IQR represents the interquartile range. All the profiles labelled as OUT_{KNN} are the upper outliers of the distribution and were removed from the set of data.

Once the abnormal load profiles were identified and filtered out, the monthly reference load profiles for each customer were calculated by averaging the remaining working daily load profiles in each month.

At this stage, in order to facilitate the subsequent grouping of similar profiles, also a normalization of data was carried out. The data normalization, especially for multidimensional problems, is necessary to compare profiles of different customers to each other removing the effect of the amplitude variability of data attributes. For energy profiling tasks, amplitude differences related to different load conditions can negatively affect the performance of pattern recognition algorithms in discovering similar shapes among profiles. To this purpose NMRLP in the (0,1) range are obtained normalizing each monthly reference load profiles respect to its maximum average power according to the following equation:

$$\hat{\boldsymbol{x}}_{i,m} = \frac{\boldsymbol{x}_{i,m}}{max(\boldsymbol{x}_{i,m})} \quad (8)$$

Where $\boldsymbol{x}_{i,m}$ is the vector representing the monthly reference load profile of the i-*th* customer for the m-*th* month and $max(\boldsymbol{x}_{i,m})$ corresponds to its maximum value.

4.2 "Follow the leader" clustering algorithm

A wide variety of clustering procedures has been introduced in the scientific literature and it is already available on different statistical softwares. The effectiveness of the different methods has been widely discussed in the literature also considering the effect of data normalization (e.g., max normalization) and data reduction (e.g., symbolic aggregate approximation, principal component analysis) techniques on the final results [60].

In this study the grouping of similar profiles was performed by means of an automatic clustering procedure based on a "Follow The Leader" (FTL) approach [37,49]. The algorithm does not require the a-priori definition of the number of clusters K, but it is initialized selecting a maximum distance threshold ρ. The dataset of the

profiles is sequentially scanned by the algorithm over a number *n* iterations, large enough to ensure the stabilization of the clustering results. In the first iteration, the FTL approach is used to define, as a first attempt, the total number of clusters K and the number of profiles assigned to each cluster. During the iterations, if the distance between a profile and the cluster centres computed until that iteration is lower than ρ*, the profile will be assigned to the cluster of the closest centroid otherwise a new cluster with one single element is generated. Indeed, the number of clusters and the number of profiles belonging to the same cluster may change until the algorithm converges to a stable solution. The algorithm was implemented in the statistical software R [61].

Given that the parameter ρ is a-priori set by the analyst, usually a cluster validity index is needed for supervising the tuning of this algorithm input. The selection of an optimal value of ρ is conducted with a "trial and error" procedure. Different values of ρ were tested and the results in terms of cluster separation and cohesion were compared through the Davies-Bouldin index.

The Davies-Bouldin Index (DBI) [62] is a cluster validity metric based on the concept that for a good partition, inter cluster separation as well as intra cluster cohesion should be as high as possible. For each clustering result obtained from the setting of different ρ, the DBI is evaluated according to the following equations:

$$DBI(\rho) = \frac{1}{K} \sum_{k=1}^{K} \max_{k \neq l} \left(\frac{\delta_k + \delta_l}{d_{k,l}} \right) \quad (9)$$

Where:

- K is the final number of clusters fixing a certain value of ρ.
- $d_{k,l}$ is the Euclidean distance between centroids of the clusters C_k and C_l.
- δ_k, δ_l are the standard deviations of the distances of objects in clusters C_k and C_l.

The value of ρ which minimises DBI is considered as the optimal value of the distance threshold for initialising the FTL algorithm.

4.3 Recursive partitioning and globally optimal classification tree

Decision trees are machine learning algorithms belonging to the family of classification/regression trees aimed at developing descriptive or predictive models from a set of records [63]. Each record can be expressed in form of tuple (x,y), where x represents the set of predictive attributes and y is the target variable. In particular, a classification tree is designed for categorical target attributes (e.g., cluster label). In this work, two classification models belonging to the family of Classification and Regression Tree (CART) were compared in order to conduct a predictive customer classification task, as it is able to easily handle both numerical and categorical target/predictive variables. CART is a binary decision tree based on the splitting of the records in purer subsets (i.e., nodes) through decision rules. The final nodes of the tree (leaves) represent the predicted class while the branches represent the conjunctions of the decision rules extracted from the predictive attributes that lead to those classes [64].

As a "baseline" classification model a decision tree was considered whose learning process is based on a recursive partitioning method. It consists of a forward step-wise approach where at each parent node the best split is evaluated maximizing homogeneity in its child nodes. However, this learning technique leads to solutions that are locally optimal, since the splits are evaluated for minimising a loss function in the next step only [64].

An alternative learning process consists of searching globally optimal trees for example by means of an evolutionary approach. This kind of classification algorithm is implemented in the *evtree* package [59] in the statistical software R. This model has been selected as "proposed" classification method.
The main steps of the algorithm can be summarised as follows [59]:

- **Setting of the model parameters**: During this step the parameters of the model were set by the analyst. The main parameters are the maximum depth of the trees, the minimum number of observations in a leaf node, the size of tree population (Θ), the variation operator probabilities, the number of iterations, the evaluation function and the complex parameter.

- **Initialization**: During this step the population of Θ trees is initialized. Each tree is initialized with a root node split randomly generated from the input variables.

- **Survivor selection**: In every iteration, each tree (parent solution) is selected once to be modified (generating an offspring solution) by one of the variation operators (i.e., split, prune, major split rule mutation, minor split rule mutation, crossover). The population size Θ remains constant during the evolution and only a fixed subset of the candidate solutions can be stored for the next iteration. The algorithm uses a deterministic crowding approach, where each parent solution competes with its most similar mutation (offspring) for being stored in the population at iteration i+1. In a classification problem the algorithm evaluates among the population of parents and offsprings, the best trees in terms of classification accuracy and complexity.

- **Termination**: The tree with the highest quality according to the evaluation function is returned as the final output of the algorithm at the end of the n iterations. For a large number of iterations (e.g. 10000 iterations) the algorithm terminates when the quality of the best 5% of trees in Θ remains stable for 100 iterations, but not before the ending of 1000 iterations.

The core of the evolutionary learning process consists in the five variation operators implemented by the algorithm at each learning iteration [59]. The main principles of the operators are described below:

1. **Split**: the operator randomly selects a leaf node of a tree and assigns a split rule to it. The split rule is randomly generated respect to the input split variable v_r and split point s_r. As a consequence, the leaf node becomes a parent node after the generation of two new child nodes;

2. **Prune**: the operator randomly selects an internal node of a tree which has two leaf nodes as successors and prunes it;

3. **Major split rule mutation**: the operator randomly chooses an internal node of a tree and modifies the split rule respect to input split variable v_r, and the split point s_r;

4. **Minor split rule mutation**: The operator randomly chooses an internal node of a tree and modifies the split rule only respect to the split point s_r of the input variable v_r;

5. **Crossover**: The operator randomly selects subtrees from two trees and exchanges them creating two new trees.

It is important to highlight that the globally optimal decision tree algorithm could lead to slightly different solutions depending on the random initialization of the population Θ and the probabilities of variation operators to be applied at each iteration. For this reason, a sensitivity analysis on the tuning of model parameters is highly recommended.
In the present paper, after the pattern recognition analysis, the labels of the most representative clusters were used as target variable in the classification process by means of a globally optimal decision tree. Different settings of tree population size, maximum number of iterations and variation operator probabilities have been tested to evaluate their impact on results.

5. Results

The methodological process presented in section 3 was implemented on the group of 114 industrial and commercial buildings. The results are presented in this section in order to demonstrate how the methodology can provide an effective tool for the automatic classification of unknown electrical energy customers.

5.1 Data pre-processing results

To perform a customer classification process, data were prepared and processed. The main objective of pre-processing phase was to evaluate the NMRLPs in a robust way. As explained in section 4.1, data pre-processing unfolds over different stages that makes it possible to automatically filter out from the dataset weekends, daily

load profiles with low standard deviation and abnormal daily load profiles. For the year 2016 the initial "*electrical power dataset*" was composed by 41.724 daily load profiles. After data pre-processing the dataset was reduced of about the 42% of the total number of daily load profiles (*Figure 6*(a)). In particular were filtered out:

- The 31% of the total amount of load profiles referred to weekends or holidays;
- The 8% of the total amount of load profiles labelled as working days that had low standard deviation of power during the day;
- The 3% of the total amount of load profiles labelled as working days that were characterized by abnormal/infrequent patterns.

The final dataset was then composed by 24.310 daily load profiles.

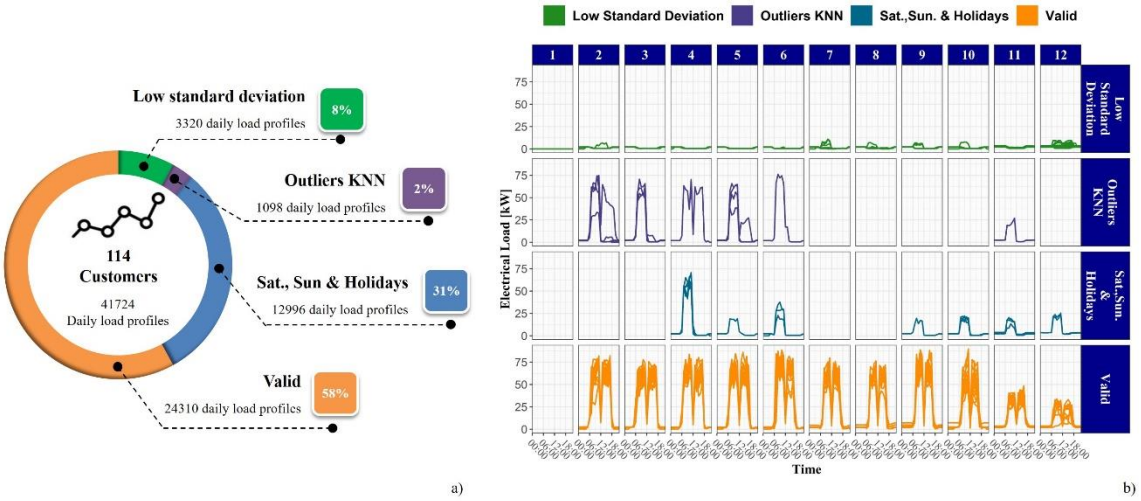

Figure 6 –Percentage of valid and excluded load profiles after pre-processing analysis (a) valid and excluded daily load profiles grouped by month for a randomly selected customer (b)

Figure 6(b) shows the impact of data pre-processing for a randomly selected customer in terms of valid and excluded daily load profiles. It is possible to notice that, after the data filtering, the remaining daily load profiles (in orange) exhibit high homogeneity in each month. This ensure that averaging those profiles per month, leads to a robust evaluation of reference patterns. At the end of the entire process the available set of 1.249 monthly reference load profiles was then normalised in the range (0,1). It is important to highlight that, although a reference period of 1 year was considered for the analysis, the number of NMRLPs per customer may be different from twelve due to the presence of missing data or the filtering of entire months during the pre-processing phase (e.g. August). On average per each customer about 10 normalized NMRLPs are available.

5.2 Clustering Results

In order to find similar groups of NMRLPs a clustering analysis has been performed. The "Follow the Leader" algorithm has been employed to this purpose as previously explained in section 4.2. The initialization of the algorithm consists in choosing an optimal value of the parameter ρ. To do this a sensitivity analysis was conducted, using the Davies Bouldin index (DBI) as reference metric for cluster validation. Considering that monthly reference load profiles have been normalized, ρ represents an a-dimensional threshold distance between load profiles in the range (0,1). The clustering results were evaluated for different values of ρ in a range between 0,8 and 2,0 with an incremental step of 0,05. For each setting of ρ the corresponding number of clusters and DBI was calculated. The results of the sensitivity analysis are shown in *Figure 7*.

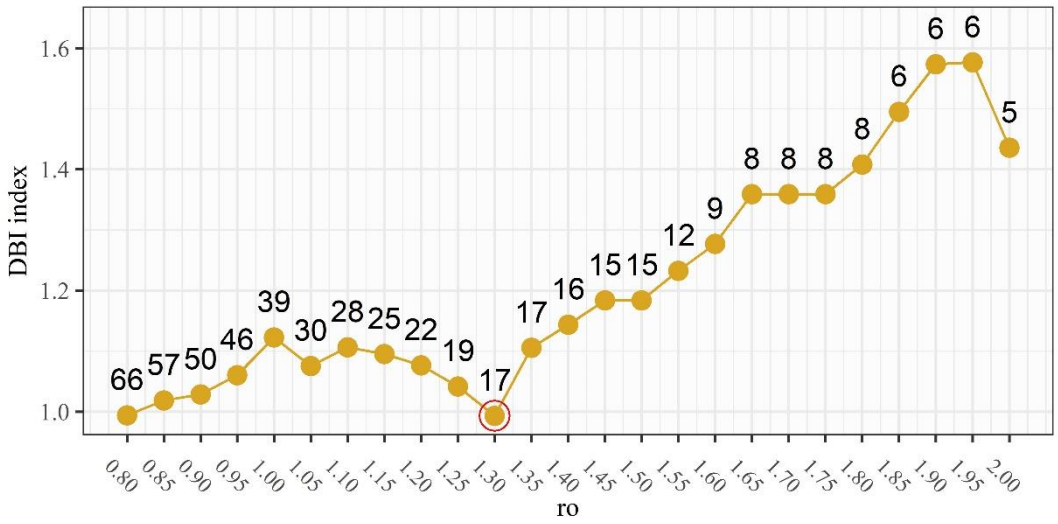

Figure 7 - Identification of optimal value ρ with the corresponding number of clusters for the initialization of "Follow the Leader" algorithm*

The *Figure 7* shows that the optimal value ρ* of the parameter ρ (that minimize the DBI) is equal to 1,30 and results in 17 clusters. It means that for ρ = ρ* the resulting clusters exhibit optimal inter cluster separation and intra cluster cohesion. The 17 clusters obtained have different cardinalities and are shown in *Figure 8* with the evidence of their centroids.

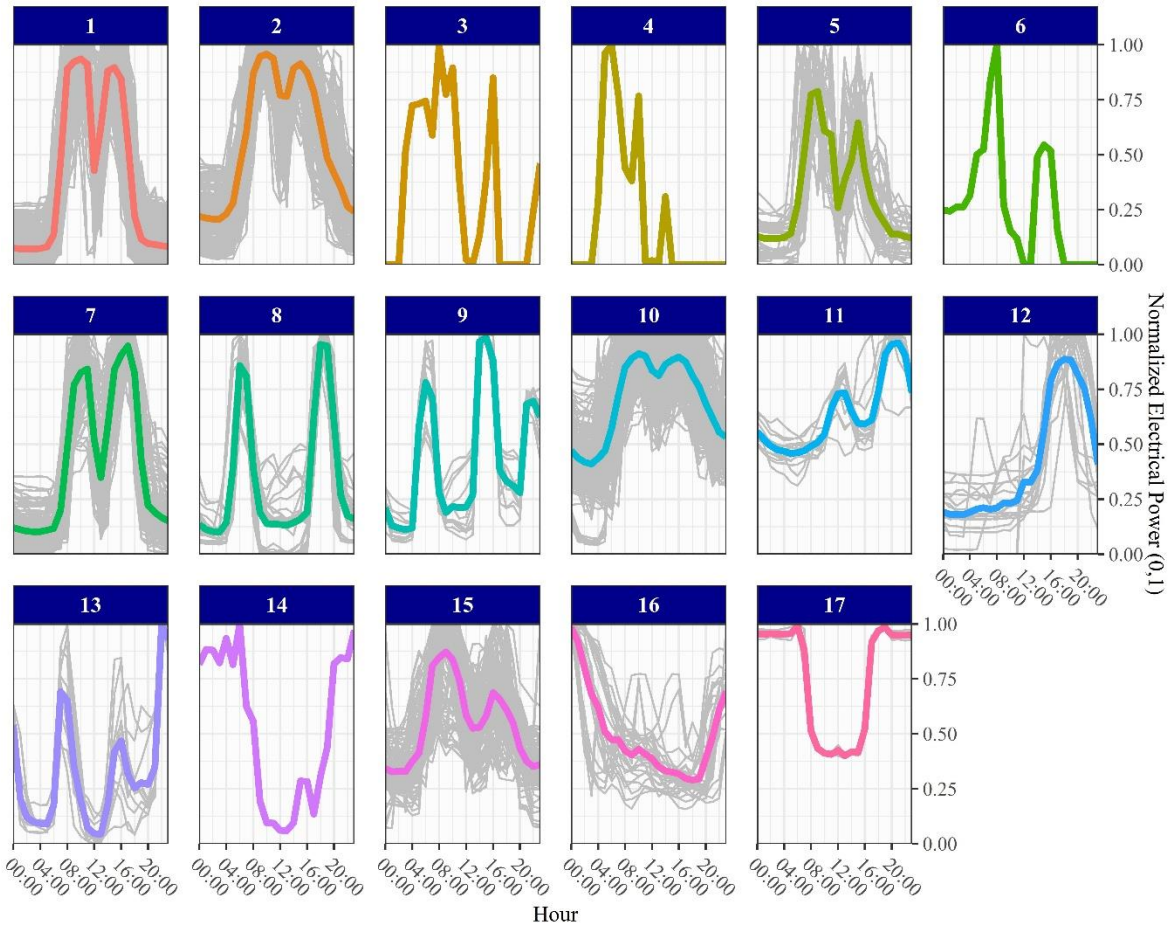

Figure 8 - Clusters of load profiles identified through the "Follow The Leader" algorithm

For classification purpose, only the most representative cluster labels were taken into account. The selection process unfolds over a descriptive analysis of the clusters obtained.

The *Figure 9* shows the scatter plot of the number of customers (x-axis) versus the number of NMRLPs (y-axis) grouped in each cluster. The horizontal and vertical dashed red lines represent the average number of NMRLPs and of customers per cluster respectively. In this way the 17 clusters are segmented according to two main space regions.

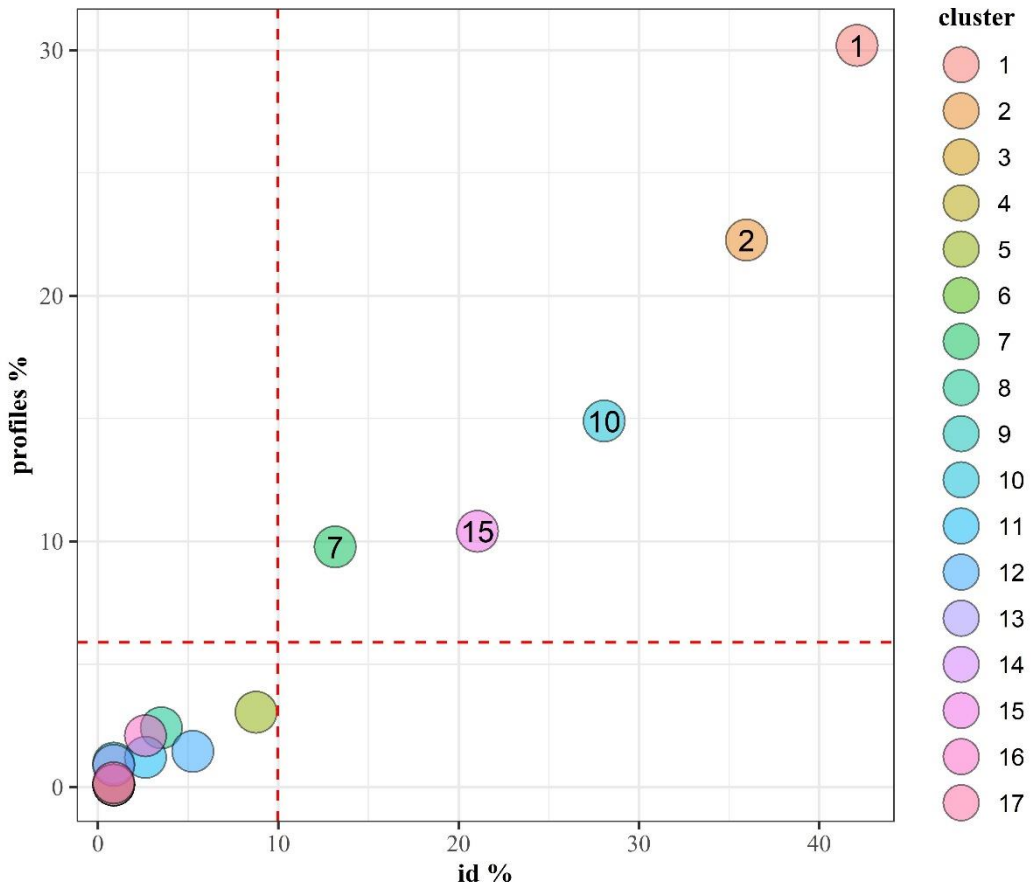

Figure 9 - Scatter plot of the number of customers (x-axis) versus the number of NMRLPs (y-axis) grouped in each cluster

The first region includes clusters in the left-bottom corner of the plot. These clusters group together few NMRLPs and customers that are characterized by patterns that significantly differ from the rest of the dataset. In particular those clusters can be described as follows:

- Clusters 3, 4, 6 and 14 include one single NMRLP. These profiles correspond to specific months during which the energy consumption patterns of some customers were infrequent. Although those profiles were not filtered out during the pre-processing phase, the “Follow-the-Leader” algorithm was able to isolate them.
- Clusters 9, 13 and 17 include all the NMRLPs of one single customer. These customers show infrequent energy consumption patterns compared to rest of the dataset and high intra cluster cohesion.
- Within Cluster 8 were grouped together customers with the same end-use which is related to milk production activities (i.e., dairy farms).
- Within clusters 11, 12, 16 were grouped together customers with end-use related to food-service activities (i.e., food industry). These are the only customers characterised by a power demand during night hours higher than in the morning ones.
- Cluster 5 includes several customers with different end-uses. However, the total number of NMRLPs that are grouped in this cluster corresponds to around the 3% of the total.

The second region includes clusters in the right-top corner of the plot. In these clusters was grouped about the 90% of the total number of NMRLPs available in the dataset corresponding to 103 out of 114 initial customers. Centroids of clusters 1, 2, 7, 10 and 15, are then generated from the most typical recognised patterns in the dataset. All these patterns are characterized by higher power demand during morning and afternoon hours than the night ones. Moreover, a reduction of power demand during the middle hours of the day occur due to the effect of a lunch-break. Although these clusters show similar trends some differences can be pointed out (see Figure 6):

- **Cluster 1** groups profiles for which power demand is high between "07:00" and "18:00" (i.e., around the 90% of the maximum power) with a strong decrease between "12:00" and "14:00" due to the lunch-break (i.e., the power demand is around the 50% of the maximum power);
- **Cluster 2** groups profiles for which the night power demand is higher than the profiles included in cluster 1 and the effect of lunch-break is less intense. Moreover, the power demand is still high also after "18:00";
- **Cluster 7** groups profiles similar to cluster 1 but for which the power demand peak occurs in the afternoon hours after the lunch-break hours;
- **Cluster 10** groups profiles with the highest power demand during night hours (i.e., the power demand is around the 30-40% of the maximum power) compared to the other clusters (i.e., cluster 1,2,7,15);
- **Cluster 15** groups profiles for which the power demand is higher in the morning hours than in the afternoon hours after the lunch-break.

In each of these clusters at least the 10% of the customers are grouped as well as about the 10% of the NMRLPs. This ensures the representativeness of such groups for customer classification purpose. For this reason, only the labels of clusters 1,2,7,10,15 were used in the subsequent phase and encoded as the categorical target variables of the decision tree. As demonstrated in other studies [37,49], the FTL algorithm has been capable to identify the most relevant clusters within the given dataset. The algorithm proved to be able to in handle outliers isolating anomalous/infrequent patterns in separate clusters that were easily identified and filtered out.

5.3 Classification Results

In this section the results obtained in the classification phase of the methodological framework are discussed. Two classification models, which are based on different learning process, were compared in terms of accuracy for predicting the cluster labels assigned to each group of NMRLPs evaluated in the clustering stage. In detail, a traditional recursive partitioning decision tree was selected as baseline, while a globally optimal decision tree was proposed as improved alternative.

Decision trees are robust and highly readable algorithm and at this stage were used to predict, for new customers, their monthly average hourly load profiles based on the membership to one of the clusters previously identified by means of FTL algorithm. It is important to notice that the prediction is monthly-based, and then a customer could have NMRLPs belonging to different clusters for each month. Therefore, in this case, the decision trees allow to finely characterise also customers with multiple typical NMRLPs among the year (e.g., presence of seasonal-based patterns).

Table 1 - Input variables for both "baseline" and "proposed" classifiers

	Description	Unit	Name
Monthly-scale Variables	Energy Consumption in time slot F1 / Total Energy consumption	-	F1
	Energy Consumption in time slot F2 / Total Energy consumption	-	F2
	Energy Consumption in time slot F3 / Total Energy consumption	-	F3
	Energy Consumption in time slot F1 / Energy Consumption in time slot F2	-	F1_2
	Energy Consumption in time slot F1 / Energy Consumption in time slot F3	-	F1_3
	Energy Consumption in time slot F2 / Energy Consumption in time slot F3	-	F2_3
Customer-features	Working start time	[h]	opening
	Working end time	[h]	closing
	Lunch break duration	[h]	d_lt

To develop the models, the input attributes were selected from the available datasets. The variables included in the model can be easily acquired through short phone survey and from customer energy bills. In this way the input data collection can be considered a non-intrusive process, since in-field energy monitoring is not needed. The input variables considered for both the "baseline" and "proposed" classifier are summarised in Table 1. All the input variables were treated as numeric or ordinal attributes, while the target variable (i.e., cluster labels) as a categorical attribute.

Before developing the classification models, from each customer cluster at least one customer was sampled (with all its NMRLPs) to be used as testing. The testing dataset consisted of 13 customers and 142 NMRLPs. Training and testing datasets were identified in order to obtain nearly the 85% of the initial population size in the training set, avoiding the presence of the NMRLPs of the same customer in both of them. Moreover, in order to roughly maintain the same share of cluster objects in the two sets, from each cluster a number of customers was sampled proportional to the cluster cardinality.

In order to perform a robust and reliable comparison, for both the "baseline" and "proposed" classifier the minimum number of elements in each leaf node (*minbucket*) and the maximum depth reachable by the tree (*maxdepth*) were set equal to 20 and 3, respectively. The *minbucket* was set equal to two times the average number of MRLPs for each customer, ensuring the presence of at least two customers classified in each leaf node of the tree. On the other hand, the maximum tree depth was set large enough to develop an accurate tree but not to much complex for avoiding overfitting issues. Considering that a *maxdepth* equal to 3 levels already limits the complexity of the possible solutions to a maximum number of 8 leaf nodes (as a consequence of three levels of binary splits), the complexity parameter α was set for both the classifiers equal to 0 in order avoid an additive penalty index in the evaluation function of the model.

Table 2 - Configurations of variation operator probabilities (globally optimal decision tree)

Setting of the variation operators	**Probabilities**				
	Crossover	**Major mutation**	**Minor mutation**	**Split**	**Prune**
c20m40sp40	20 %	20 %	20 %	20 %	20 %
c10m30sp60	10 %	15 %	15 %	30 %	30 %
c00m50sp50	-	25 %	25 %	25 %	25 %
c40m20sp40	40 %	10 %	10 %	20 %	20 %
c10m10sp80	10 %	5 %	5 %	40 %	40 %
c50m00sp50	50 %	-	-	25 %	25 %

For the "proposed" decision tree based on the evolutionary learning algorithm, further hyper parameters need to be set. The parameters to be tuned are the population size Θ, the maximum number of iterations and the variation operator probabilities. Six different configurations of variation operator probabilities, three different number of maximum iterations and four population sizes have been tested. This analysis unfolds over two steps, as presented in [59].

In the first step, the 18 configurations generated by combining six different settings of variation operator probabilities (*Table 2*) and three maximum number of iterations (i.e., 500, 1000, 10000) have been analysed. Each combination has been tested for 100 different random initialisations of the population Θ, which is fixed at 100 trees (default value). Each solution was evaluated computing its misclassification error. *Figure 10* shows the box plots of the 100 misclassification errors for each of the 18 combinations of the trees developed on the entire dataset.

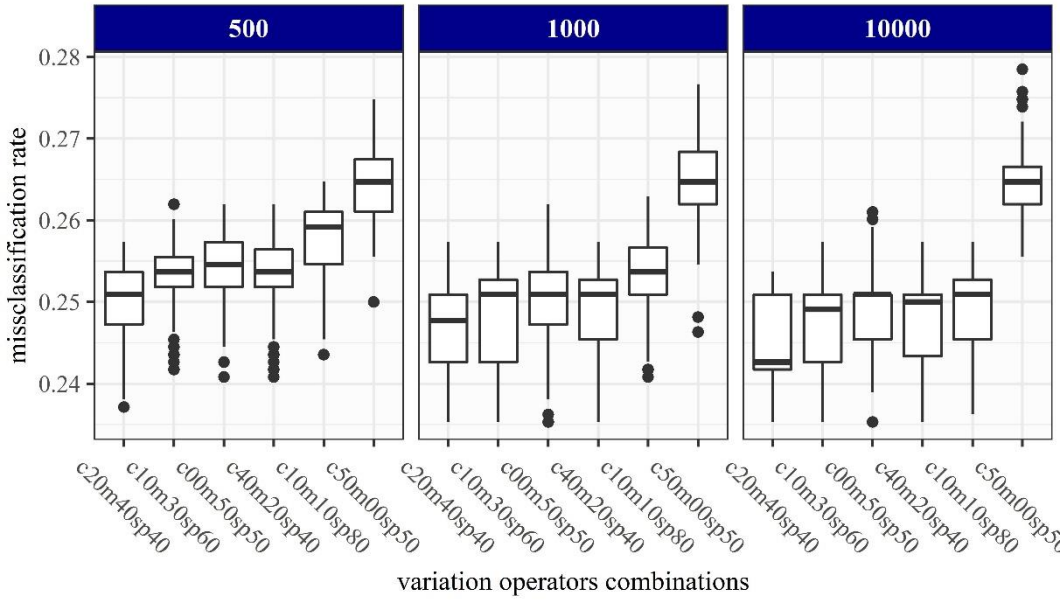

Figure 10 - Misclassification rates for 18 configurations of iteration number and variation operator probabilities (globally optimal decision tree)

From this first step it is possible to infer that the misclassification rate of the decision trees decreases with the increasing of the maximum number of iterations reaching its best median value for 10000 iterations and variation operator probabilities set at *c20m40sp40.* In the second step of the analysis the impact of the population size Θ on the misclassification rate is evaluated considering 100 different random initialisations of populations with size of 25, 50, 100, 250 and 500 trees respectively. In this step the number of iterations and variation operator probabilities were set at 10000 and *c20m40sp40* respectively, that correspond to the optimal values previously identified.

Figure 11 shows that the the cardinality of population size positively affects the overall performance of the decision tree, reaching the minimum median value of the misclassification rate for a population of 500 trees.

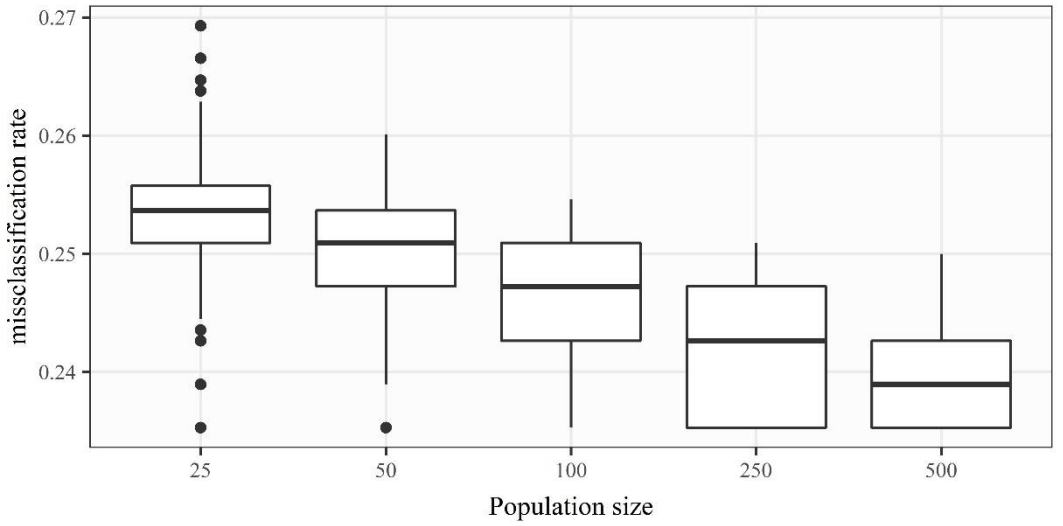

Figure 11 - Misclassification rates for populations with size of 25, 50, 100, 250 and 500 trees respectively (globally optimal decision tree)

According to the performed sensitivity analysis, the globally optimal tree was then developed on the training dataset with the following parameter setting: Θ equal to 500, number of iterations set to 10000 and the variation operator probabilities set to *c20m40sp40.*

The final decision trees (i.e., “baseline” vs “proposed”), developed on the training dataset are shown in *Figure 12* and *Figure 13* respctevely. The two trees differ in terms of number of leaf nodes and input variables used for the split generation. The globally optimal decision tree was capable to converge into a more detailed and accurate solution following decision rule paths different from the other model.

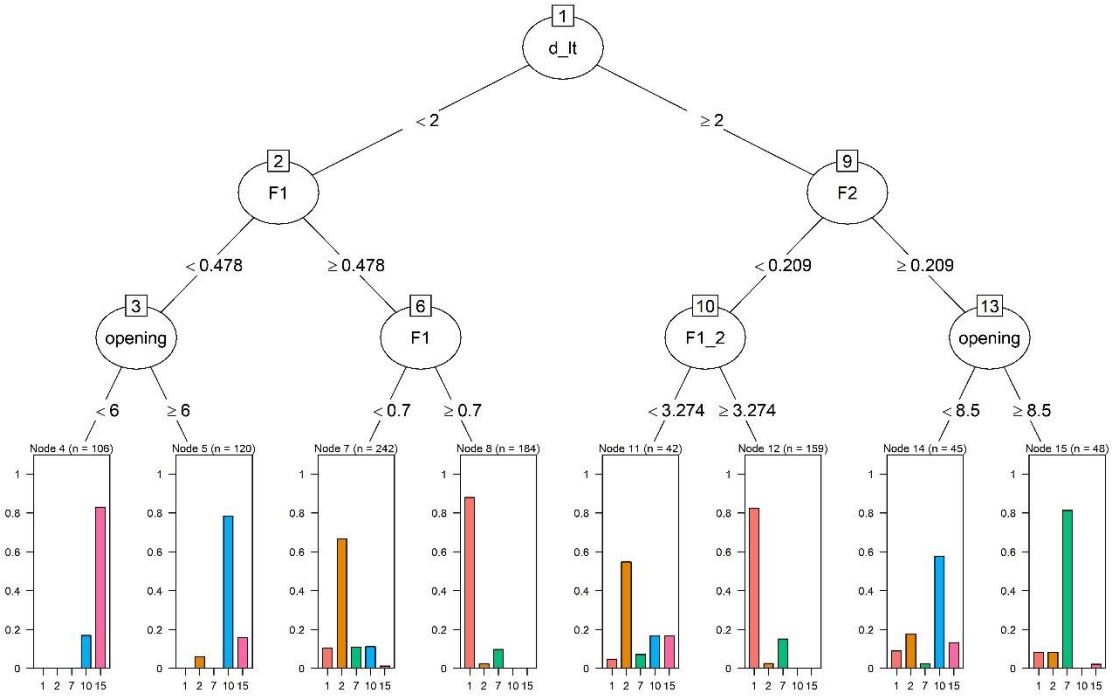

Figure 12 - Globally optimal decision tree

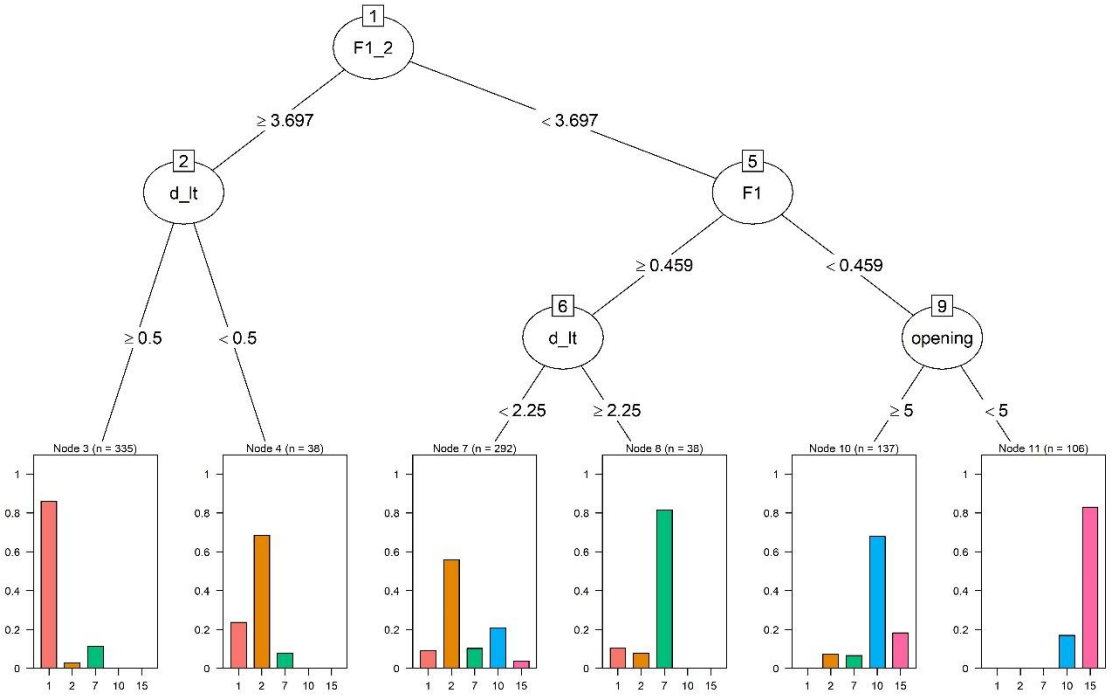

Figure 13 – Recursive partitioning decision tree

In fact, the locally optimal decision tree at each parent node evaluates the best split, maximizing homogeneity in the next step only. On the contrary, the globally optimal decision tree is capable to leverage on less accurate internal splits in order to reach a higher final performance of the classifier. In *Table 3* and *Table 4* the decision rules extracted from the two classifiers are reported.

Table 3 - Decision rules extracted from globally optimal classifier

Cluster	Node	Decision Rules	Profiles	Accuracy
1	8	IF **d_lt** < 2 AND **F1** ≥ 0.504 AND **F1** ≥ 0.701	184	88 %
	12	IF **d_lt** ≥ 2 AND **F2** < 0.208 AND **F1_2** ≥ 3.27	159	82.4%
2	7	IF **d_lt** < 2 AND **F1** ≥ 0.504 AND **F1** < 0.701	230	67.8 %
	11	IF **d_lt** ≥ 2 AND **F2** < 0.208 AND **F1_2** < 3.27	42	54.8 %
7	15	IF **d_lt** ≥ 2 AND **F2** ≥ 0.208 AND **opening** ≥ 08:30	48	81.7 %
10	5	IF **d_lt** < 2 AND **F1** < 0.504 AND **opening** ≥ 06:00	131	75.6 %
	14	IF **d_lt** ≥ 2 AND **F2** ≥ 0.208 AND **opening** < 08:30	45	57.8 %
15	4	IF **d_lt** < 2 AND **F1** < 0.504 AND **opening** < 06:00	107	82.2 %

Table 4 - Decision rules extracted from recursive partitioning classifier

Cluster	Node	Decision Rules	Profiles	Accuracy
1	3	IF **F1_2** ≥ 3.697 AND **d_lt** ≥ 0.5	335	86 %
2	4	IF **F1_2** ≥ 3.697 AND **d_lt** < 0.5	38	68.4 %
	7	IF **F1_2** < 3.697 AND **F1** ≥ 0.459 AND **d_lt** < 2.25	292	55.8 %
7	8	IF **F1_2** < 3.697 AND **F1** ≥ 0.459 AND **d_lt** ≥ 2.25	38	81.6 %
10	10	IF **F1_2** < 3.697 AND **F1** < 0.459 AND **opening** ≥ 05:00	137	67,9 %
15	11	IF **F1_2** < 3.697 AND **F1** < 0.459 AND **opening** < 05:00	106	83 %

Both models suggest that NMRLPs grouped within cluster 1 and cluster 2 are characterised by a higher monthly energy consumption during time slot F1 respect to other clusters. However, the energy consumption during F2 hours are more significant for cluster 2 compared to cluster 1. According to the globally optimal decision tree solution, customers whose NMRLPs were grouped within cluster 7 are characterized by working activities starting later than 08:30 a.m., while such a feature is not extractable from baseline solution (recursive partitioning decision tree). Whithin cluster 10 and cluster 15, were grouped NMRLPs for which the energy consumption during time slot F2 are higher compared to other clusters. The difference between the cluster 10 and 15 consists in an earlier starting of working activities for customers in cluster 15 than of others in cluster 10. Those cluster features have been exploited by both models, however, the globally optimal decision tree showed a more detailed description by using one more decision rule. The rules are furtherly applied on the testing set to evaluate “out-of-sample” performances of the two models.

Table 5 - Overall misclassification errors of recursive partitioning and globally optimal decision tree for the training and testing dataset

	Misclassification error	
	Globally optimal decision tree	***Recursive partitioning decision tree***
training	23.5 %	27.1 %
testing	24.6 %	30.9 %

In

the overall misclassification errors of the two models are reported for the training and testing datasets. It is possible to see that the proposed globally optimal decision tree performs better than the locally optimal one both in training and testing. The accuracy in testing session improves by about 6%.

Although for the proposed model the setting of parameters is not straightforward and the computational cost is quite high, the algorithm is capable to reach results significantlly better than the baseline approach in terms of generalizability and accuracy of the model.

5.4 Rescaling process

The last phase of the methodological process consists of the rescaling of the estimated NMRLPs. Infact, after the classification of the 13 customers of the testing dataset, their estimated NMRLPs were rescaled in order to obtain a reference hourly power demand profile expressed in kW. The NMRLPs were rescaled by multipling their 24 values (one for each hour of the day) by the scaling factor K (as explained in section 3) obtained by using the actual monthly energy consumption in the F1 time slot. The rescaled NMRLPs were compared to the actual ones in order to evaluate the overall performance of the methodological framework. In particular the Pearson correlation computed between real and rescaled profiles was select as validity index. For instance, in the case of a customer with 10 monthly reference load profiles, the correlation coefficient is calculated among 24 x 10 data points expressed in kW. On average for the entire testing set, consisting in 142 profiles, a strong linear correlation coefficient equal to *0.895* was obtained (*Figure 14*). It means that the process is capable to return, for a unknown customer, a set of estimated monthly reference load profiles that are accurate in terms of both magnitude and shape.

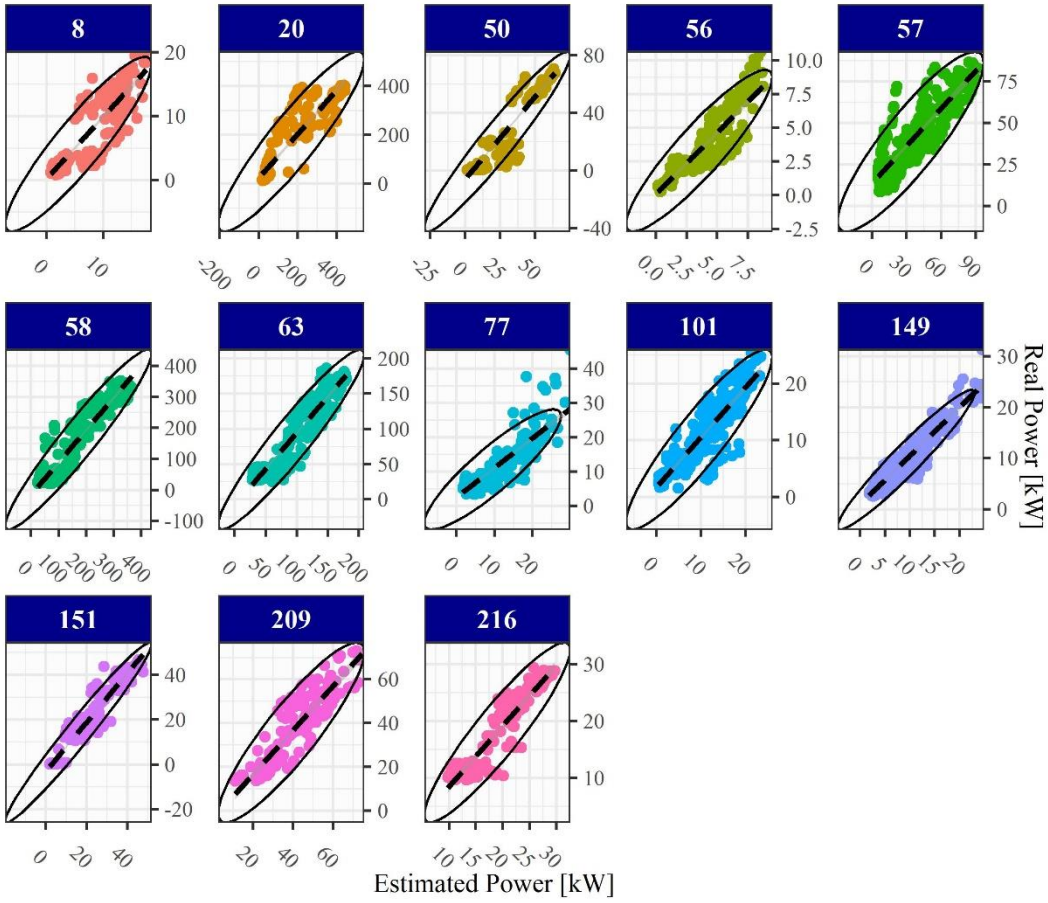

Figure 14 - Linear correlations between actual and rescaled estimated energy profiles for each customer of the testing set

As a reference, in *Figure 15* (a) the results of the rescaling process are shown for each month of a randomly selected customer from the testing set. For each month, grey lines show the actual load profiles of the working days, the red line is the actual average profile and blue line is the rescaled NMRLP. In addition in *Figure 15* (b) the carpet plot of actual load profiles is reported together with carpet plot reconstructed on monthly basis through the rescaled estimated load profiles. Both figures show how the process performs proving its robustness and effectiveness.

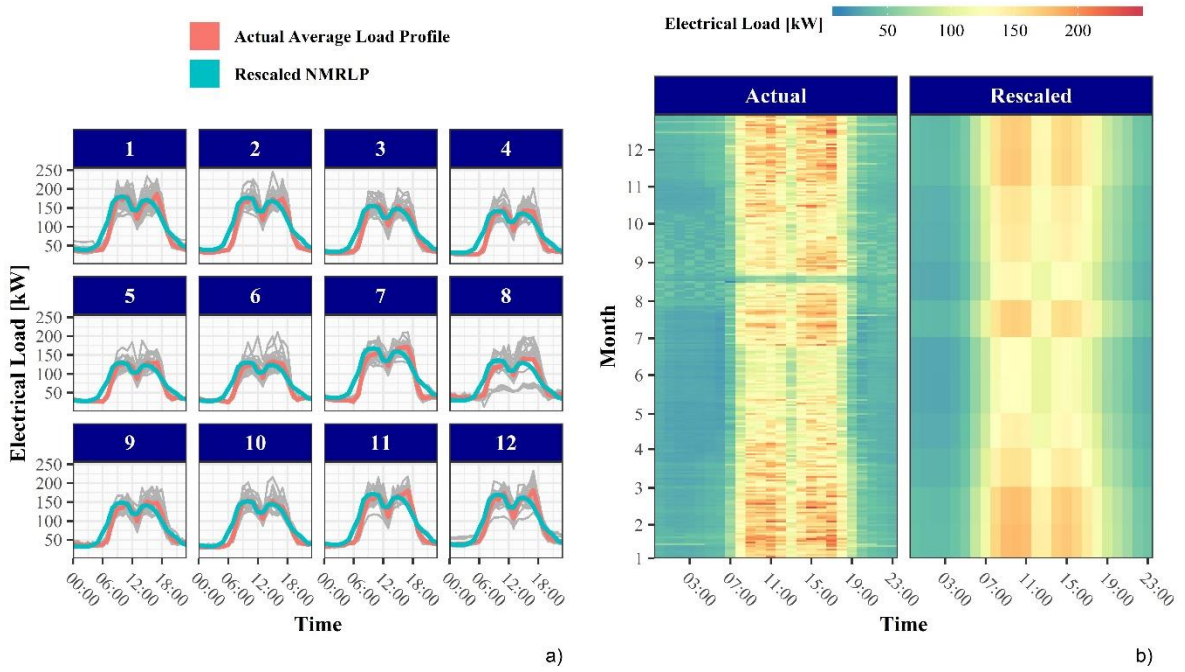

Figure 15 – Actual load profiles of the working days (grey lines), actual average load profiles (red lines) and rescaled load profiles (blue lines) of a randomly selected customer from the testing set (a) carpet plot of actual load profiles together with the carpet plot reconstructed on monthly basis through the rescaled estimated load profiles (b)

6. Discussion

The present paper focused on the analysis of electrical load patterns of a stock of 114 industrial and commercial buildings located in Piedmont (North-Western region of Italy). The proposed methodology provides a robust process for the automatic classification of unknown electrical customers.

For this purpose, on the basis of the existing literature, proven algorithms were employed in the analysis. In the pattern recognition phase the "follow the leader" approach has been used for identifying the most significant customer groups and at the same time isolate infrequent or anomalous patterns in separate groups. The algorithm belongs to the family of partitional clustering techniques, but differently from K-means it requires a distance threshold instead of the number of desired clusters K as input parameter. It brings advantages in terms of algorithm flexibility. In fact the use of a distance threshold makes it possible to better manage infrequent/anomalous patterns without previously perform an outlier removal analysis for improving clustering performances. The setting of the threshold ρ has been supervised by using the Davies Bouldin Index as a cluster validity metric allowing the optimal value ρ* to be automatically identified. The cluster analysis resulted in 17 customer groups characterized by different cardinality and shapes. Even if the high diversity of patterns represents an asset in a customer classification process, a large customer database is required to adequately represent each of them.

In the presented case study, 5 cluster labels were used in the classification phase given that the remaining 12 groups included few customer profiles or anomalous ones. Excluding Clusters 3, 4, 6 and 14 that included one single anomalous NMRLP, the others are candidates for being considered in the classification process when further NMRLPs will be stored in the customer database. In that perspective, the process can be considered open and furtherly upgradable considering that more cluster labels could be taken into account in the future for developing an extended classifier. One of the most recently developed algorithm for decision tree based on globally optimal learning process was tested and compared with the well-known one-step-forward approach. This proposed classifiction model led to an improved accuracy of 6% for the testing data set in comparison to the baseline classification model. Differently from more straightforward decision tree models, the globally optimal algorithm requires a high computational cost and the tuning of model parameters represents a time consuming task. For the case analysed in this work, the higher accuracy achieved and the limited database volume made its implementation still reasonable. The algorithm was capable to accomplish the classification

task by fully exploiting the few input variables collected through a non-intrusive approach. In the authors opinion, this hypothesis represents one of the strengths of the methodological process proposed, given that it allows users to preliminary characterize electric or thermal energy customers in a very detailed way without using in-field monitoring data [65]. The opportunity to estimate, for an unknown customer, its most probable NMRLPs is highly desirable for several stakeholders (e.g., suppliers, local and national authorities) in the smart city environment.
As a consequence, more effective energy management strategies can be conceived for different customer groups on the basis of their representative load patterns for example by designing targeted financial demand response programs (e.g., Time Of Use tariff, Critical Peak Pricing, Real-Time Pricing).
These programmes are getting more and more attention as retailes keep looking for a better way to balance loads and at the same time increase their profitability. On the other hand, such programs are designed to be attractive also for the consumers as they can exploit a deeper knowledge of their energy patterns to reduce the total energy bill cost. In this context the modification of a load profile plays a critical role not only from an economical point of view but also in terms of grid stability.
Customer classification tools can also be employed for traking the changes of power consumption patterns over time. By benchmarking customer flexibility (in terms of demand modification) it is possible to assess which could be the influence and the impact of specific Demand Side Management and Demand Response initiatives for a group of customers or even at larger scale (e.g. district). Differently from the literature, the classification process, adopted in the work presented here, is also capable to estimate, the magnitude of energy profiles. The results proved that the methodological process introduced allows to robustly estimate for a unknown customer, a set of monthly reference load profiles that are accurate in terms of both magnitude and shape.
The opportunity to estimate the shape of a load profile thogheter with its magnitude enables a full characterization of a building energy demand, making it possible to easily and effectively reach decarbonisation targets also from system design side.

Conclusions

In the present paper, the application of a non-intrusive approach for addressing a customer classification task was investigated. The analysed stock of customers consisted in more than 100 non-residential buildings with 17 different end-use categories. A classification tool capable to predict for a new customer its most probable typical load monthly profiles was developed. The classification model makes use of an globally optimal decision tree algorithm that differently from traditional recursive partitioning decision tree leverages on an evolutionary learning process in searching optimal decision trees. The model was fed by predictive attributes extracted from monthly energy bills of each customer and from additional information collected by means of phone survey. Despite the use of non-shape sensitive attributes, the model reached an overall accuracy of about 80%. The conceived procedure makes it possible to exploit energy bill data also for estimating the magnitude of typical load profiles.
Eventually the proposed non-intrusive methodology showed good performance significantly improving the feasibility of a customer classification process in real-life applications.

Acknowledgements

This study was supported by eVISO s.r.l (www.eviso.it) in the framework of a research contract with Department of Energy of Politecnico di Torino. The authors would like to express their gratitude to Eng. Carlo Cigna and Eng. Gianfranco Sorasio for the suggestions and the support in the research activity.

References

[1] Tureczek A, Nielsen PS, Madsen H. Electricity consumption clustering using smart meter data. Energies 2018; 11: 1–18.

[2] Pérez-Chacón R, Luna-Romera JM, Troncoso A, Martínez-Alvarez F, Riquelme JC. Big data analytics for discovering electricity consumption patterns in smart cities. Energies 2018; 11: 1–19.

[3] Capozzoli A, Piscitelli MS, Brandi S. Mining typical load profiles in buildings to support energy management in the smart city context. Energy Procedia 2017; 134: 865–874.

[4] Liu X. Smart Meter Data Analytics: Systems, Algorithms, and Benchmarking. ACM Trans Database Syst 2016; 42: 1–39.

[5] Miller C, Nagy Z, Schlueter A. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. Renew Sustain Energy Rev 2016; 81: 1365-1377

[6] Fan C, Xiao F, Li Z, Wang J. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy Build 2018; 159: 296–308.

[7] Fan C, Xiao F, Wang S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energy 2014; 127: 1–10.

[8] Yu Z, Haghighat F, Fung BCM, Zhou L. A novel methodology for knowledge discovery through mining associations between building operational data. Energy Build 2012; 47: 430–440.

[9] Kim W, Katipamula S. A review of fault detection and diagnostics methods for building systems. Sci Technol Built Environ 2018; 24: 3–21.

[10] Capozzoli A, Piscitelli MS, Gorrino A, Ballarini I, Corrado V. Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. Sustain Cities Soc 2017; 35: 191-208

[11] Yu Z, Fung BCM, Haghighat F, Yoshino H, Morofsky E. A systematic procedure to study the influence of occupant behavior on building energy consumption. Energy Build 2011; 43: 1409–1417.

[12] Capozzoli A, Piscitelli MS, Brandi S, Grassi D, Chicco G. Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. Energy 2018; 157: 336–352.

[13] Iglesias F, Kastner W. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. Energies 2013; 6: 579–597.

[14] Miller C, Meggers F. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. Energy Build 2017; 156: 360–373.

[15] Park JY, Yang X, Miller C, Arjunan P, Nagy Z. Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. Appl Energy 2019; 236:

[16] Luo X, Hong T, Chen Y, Piette MA. Electric load shape benchmarking for small- and medium-sized commercial buildings. Appl Energy 2017; 204: 715–725.

[17] Wang Y, Chen Q, Kang C, Zhang M, Wang K, Zhao Y. Load profiling and its application to demand response: A review. Tsinghua Sci Technol 2015; 20: 117–129.

[18] Arco L, Casas G, Nowé A. Clustering methodology for smart metering data based on local and global features, in: IML '17 Proc. 1st Int. Conf. Internet Things Mach. Learn., 2017; 1–13.

[19] Panapakidis IP, Christoforidis GC. Implementation of modified versions of the K-means algorithm in power load curves profiling. Sustain Cities Soc 2017; 35: 83–93.

[20] Zakovorotnyi A, Seerig A. Building energy data analysis by clustering measured daily profiles. Energy Procedia 2017; 122: 583–588.

[21] Tureczek AM, Nielsen PS. Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data. Energies 2017; 10: 1–19.

[22] Dudek G. Neural networks for pattern-based short-term load forecasting: A comparative study. Neurocomputing 2016; 205: 64–74.

[23] Fan C, Xiao F, Yan C. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Autom Constr 2015; 50: 81–90.

[24] Miller C, Nagy Z, Schlueter A. Automated daily pattern filtering of measured building performance data. Autom Constr 2015; 49: 1–17.

[25] Capozzoli A, Piscitelli MS, Brandi S, Grassi D, Chicco G. Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. Energy 2018; 157: 336–352.

[26] Panapakidis IP, Papadopoulos TA, Christoforidis GC, Papagiannis GK. Pattern recognition algorithms for electricity load curve analysis of buildings. Energy Build 2014; 73: 137–145.

[27] Do Carmo CMR, Christensen TH. Cluster analysis of residential heat load profiles and the role of technical and household characteristics. Energy Build 2016; 125: 171–180.

[28] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Appl Energy 2014; 135: 461–471.

[29] Wang F, Zhen Z, Wang B, Mi Z, Wang Z, Li K. A Baseline Load Estimation Approach for Residential Customer based on Load Pattern Clustering. Energy Procedia 2018; 142: 2042–2049.

[30] Grigoraș G, Bobric E-C. Clustering Based Approach for Customers' Classification From Electrical Distribution Systems. UPB Sci Bull, Ser C 2015; 77: 219–226.

[31] Siano P. Demand response and smart grids - A survey. Renew Sustain Energy Rev 2014; 30: 461–478.

[32] Gelazanskas L, Gamage KAA. Demand side management in smart grid: A review and proposals for

future direction. Sustain Cities Soc 2014; 11: 22–30.

[33] Verda V, Guelpa E, Sciacovelli A, Acquaviva A, Patti E. Thermal peak load shaving through users request variations. Int J Thermodyn 2016; 19: 168–176.

[34] Jang D, Eom J, Jae Park M, Jeung Rho J. Variability of electricity load patterns and its effect on demand response: A critical peak pricing experiment on Korean commercial and industrial customers. Energy Policy 2016; 88: 11–26.

[35] Chen CS, Hwang JC, Huang CW. Application of load survey systems to proper tariff design. IEEE Trans Power Syst 1997; 12: 1746–1751.

[36] Wang K, Zhang M, Wang Z, Li R, Li F, Wu H. Time of use tariff design for domestic customers from flat rate by model-based clustering. Energy Procedia 2014; 61: 652–655.

[37] Chicco G, Napoli R, Postolache P, Scutariu M, Toader C. Customer Characterization Options for Improving the Tariff Offer. IEEE Trans Power Syst 2003; 18: 381–387.

[38] Capozzoli A, Cerquitelli T, Piscitelli MS. Chapter 11 – Enhancing energy efficiency in buildings through innovative data analytics technologies, in: D. Ciprian, F. Xhafa (Eds.), Pervasive Comput., 2016: pp. 353–389.

[39] Jalali MM, Kazemi A. Demand side management in a smart grid with multiple electricity suppliers. Energy 2015; 81: 766–776.

[40] Azaza M, Wallin F. Smart meter data clustering using consumption indicators: Responsibility factor and consumption variability. Energy Procedia 2017; 142: 2236–2242.

[41] Benítez I, Quijano A, Díez JL, Delgado I. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. Int J Electr Power Energy Syst 2014; 55: 437–448.

[42] Khan I, Huang JZ, Luo Z, Masud MA. CPLP: An algorithm for tracking the changes of power consumption patterns in load profile data over time. Inf Sci 2018; 429: 332–348.

[43] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 2012; 42: 68–80.

[44] Panapakidis I, Christoforidis G. Optimal Selection of Clustering Algorithm via Multi-Criteria Decision Analysis (MCDA) for Load Profiling Applications. Appl Sci 2018; 8: 237–279.

[45] Panapakidis I, Alexiadis M, Papagiannis G. Evaluation of the performance of clustering algorithms for a high voltage industrial consumer. Eng Appl Artif Intell 2015; 38: 1–13.

[46] Tsekouras GJ, Hatziargyriou ND, Dialynas EN. Two-stage pattern recognition of load curves for classification of electricity customers. IEEE Trans Power Syst 2007; 22: 1120–1128.

[47] Mcloughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile

characterisation using smart metering data. Appl Energy 2015; 141: 190–199.

[48] Fernandes MP, Viegas JL, Vieira SM, Sousa JMC. Segmentation of residential gas consumers using clustering analysis. Energies 2017; 10: 2047–2073.

[49] Chicco G, Napoli R, Postolache P, Scutariu M, Toader C. Emergent electricity customer classificatio. IEE Proceedings-Generation, Transm Distrib 2005; 152: 164–172.

[50] Figueiredo V, Rodrigues F, Vale Z, Gouveia JB. An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques. IEEE Trans Power Syst 2005; 20: 596–602.

[51] Chicco G, Ilie IS. Support vector clustering of electrical load pattern data. IEEE Trans Power Syst 2009; 24: 1619–1628.

[52] Yang J, Ning C, Deb C, Zhang F, Cheong D, Eang Lee S, Sekhar C, Wai Tham K. k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy Build 2017; 146: 27–37.

[53] Ma Z, Yan R, Nord N. A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings. Energy 2017; 134: 90–102.

[54] Piao M, Ryu KH. Subspace Frequency Analysis-Based Field Indices Extraction for Electricity Customer Classification. ACM Trans Inf Syst 2016; 34: 1–18.

[55] Biscarri F, Monedero I, García A, Guerrero JI, León C. Electricity clustering framework for automatic classification of customer loads. Expert Syst Appl 2017; 86: 54–63.

[56] Zhong S, Tam KS. Hierarchical Classification of Load Profiles Based on Their Characteristic Attributes in Frequency Domain. IEEE Trans Power Syst 2015; 30: 2434–2441.

[57] Ramos S, Duarte JM, Duarte FJ, Vale Z. A data-mining-based methodology to support MV electricity customers' characterization. Energy Build 2015; 91: 16–25.

[58] Bicego M, Farinelli A, Grosso E, Paolini D, Ramchurn SD. On the distinctiveness of the electricity load profile. Pattern Recognit 2018; 74: 317–325.

[59] Grubinger T, Zeileis A, Pfeiffer K-P. evtree : Evolutionary Learning of Globally Optimal Classification and Regression Trees in R. J Stat Softw 2015; 61: 1–29.

[60] Notaristefano A, Chicco G, Piglione F. Data size reduction with symbolic aggregate approximation for electrical load pattern grouping. IET Gener Transm Distrib 2013; 7: 108–117.

[61] R Core Team. R: A Language and Environment for Statistical Computing. 2017; http://www.r-project.org/.

[62] Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell 1979; PAMI-1: 224–227.

[63] Tan P-N, Steinbach M, Kumar V. Classification: Basic Concepts, Decision Trees, and Model Evaluation. Introd to Data Min 2006; 67: 145–205.

[64] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees, 1984.

[65] Vercamer D, Steurtewagen B, Van Den Poel D, Vermeulen F. Predicting Consumer Load Profiles Using Commercial and Open Data. IEEE Trans Power Syst 2016; 31: 3693–3701.